Research Report

Examining the Impact of Audio Presentation on Tests of Reading Comprehension

Cara Cahalan Laitusis

Linda L. Cook

Frederick Cline

Teresa C. King

John Sabatini

April 2008

ETS RR-08-23

Examining the Impact of Audio Presentation on Tests of Reading Comprehension

Cara Cahalan Laitusis, Linda L. Cook, Frederick Cline, Teresa C. King, and John Sabatini
ETS, Princeton, NJ

April 2008

Abstract

This study examined the impact of a read-aloud accommodation on standardized test scores of reading comprehension at Grades 4 and 8. Under a repeated measures design, students with and without reading-based learning disabilities took both a standard administration and a read-aloud administration of a reading comprehension test. Results show that the mean score on the audio version was higher than scores on the standard version for both groups of students at both grade levels. Students with reading-based learning disabilities at both levels benefited differentially more than students with no disability. This finding continues to hold after controlling for reading fluency and ceiling effects at both grades. The results also examined the relationship between test scores and teachers' ratings of reading comprehension to determine which measures are the best predictors of teachers' ratings of reading comprehension by grade and disability classification.

Key words: Reading, learning disabilities, accommodations, read aloud, NCLB, modifications, validity

Acknowledgments

Funding was provided by Grant No. H324F040001 by the U.S. Department of Education Office of Special Education and Rehabilitation Services (OSERS) and the Institute of Education Sciences (IES) National Center of Special Education Research (NCSER).

Table of Contents

Page

Introduction 1

Review of Research 3

Limitations of Prior Research 5

Research Questions 6

Method 6

Sample 6

Selection of Schools 6

Selection of Students 6

Description of Final Sample 7

Materials 8

Assessments 9

Surveys 10

Procedure 11

Results 12

Differential Boost 12

Differential Boost Controlling for Fluency 15

Differential Boost Controlling for Ceiling Effect 16

Predictive Validity of Audio and Standard Scores 17

Correlations 18

Regression Analyses 22

Conclusions 25

Limitations 26

Future Research 27

References 28

List of Appendixes 31

List of Tables

Page

Table 1. Percentage of Students by Race, Grade, and Disability Category 7

Table 2. Number and Percent of Reading-Based Learning Disabilities by Aspect of Reading Impacted by Disability 8

Table 3. Design for Gates-McGinitie Test Administration... 11

Table 4. Sample Size by Experimental Group, Grade, and Disability Status........................... 12

Table 5. Means and Standard Deviations for Scaled Scores Gates-McGinitie Reading Test (GMRT; Standard, Audio, and Boost) and Woodcock-Johnson III Diagnostic Reading Battery (WJ-III DRB) Reading Fluency Raw Score by Grade and Disability Status 13

Table 6. Repeated Measures Analysis of Variance (RM-ANOVA) for Grade 4 14

Table 7. Repeated Measures Analysis of Variance (RM-ANOVA) for Grade 8 15

Table 8. Repeated Measures Analysis of Covariance (RM-ANCOVA) for Grade 4 With Fluency as a Covariate 16

Table 9. Repeated Measures Analysis of Covariance (RM-ANCOVA) for Grade 8 With Fluency as a Covariate 17

Table 10. Intercorrelations Between Test Scores, Boost, and Teachers' Ratings for Grade 4... 19

Table 11. Intercorrelations Between Test Scores, Boost, and Teachers' Ratings for Grade 8... 20

Table 12. Comparison of Alternate Measurement Models to Model 1—Standard Gates-McGinitie Reading Test (GMRT).. 23

Introduction

Recent legislation, such as the Individuals with Disabilities Education Act (IDEA; IDEA, 1997; IDEA, 2004) and the No Child Left Behind Act (NCLB; NCLB, 2001), has increased the participation of students with disabilities in statewide achievement testing. Prior to 1997, students in special education were often excluded from this type of testing. The reauthorization of IDEA '97 mandated that students with disabilities be included in standardized assessments and that accommodations be made where appropriate for their inclusion. IDEA further clarified that states had to provide accommodation guidelines and report on the number of students using accommodations. In 2001, NCLB redefined the federal government's role in K–12 education. Along with mandating annual student testing in Grades 3 to 8, the act stipulates that students with disabilities receive test accommodations as defined in the Americans With Disabilities Act of 1990 (ADA, 1990) and IDEA.

Most states differentiate between testing accommodations and testing modifications and provide a list of each in their guidelines for testing students with disabilities and English language learners. While accommodations are changes to testing procedures or materials that do not alter the construct being assessed or the comparability of test scores (between accommodated and nonaccommodated conditions), testing modifications do alter the construct being tested and consequently affect the comparability of test scores. Modifications are sometimes referred to as nonstandard administrations or nonallowable accommodations (Thurlow & Wiener, 2000). A recent review of state policy on testing accommodations found that the vast majority of states consider most test changes to be testing accommodations (Clapper, Morse, Lazarus, Thompson, & Thurlow, 2005). For example, nearly all states agree that extra time is an accommodation (not a modification) on state assessments.

States are not in agreement, however, on whether to consider the audio presentation of test content (i.e., read aloud) on reading assessments to be an accommodation or a modification. These differences are largely due to different specifications of reading in each state's reading standards. Some states (e.g., California and New Jersey) have determined that reading involves visual or tactile decoding of text, while others (e.g., Wisconsin) argue that when a reading test is read aloud the "nature of what the test is measuring (reading comprehension) has been changed to one of listening comprehension" (Wisconsin Department of Public Instruction, 2003). States that allow read-aloud accommodations on tests of reading or English language arts (e.g., Kentucky and Delaware) have defined reading as comprehension of written material that is presented in a visual, tactile, or audio format. Even states that consider read-aloud a modification, significant numbers of students participate in testing with that modification. In California, for example, over 5,000 fourth grade students (representing nearly 11% of students in special education) took the Standardized Testing and Reporting (STAR) English Language Arts assessment in 2006 as a read-aloud test (ETS, 2007). In sum, states are struggling to present reading assessments that are accessible to students with reading-based learning disabilities (RLDs) yet also provide valid measures of the construct of reading.

It is clear that if the assessment is designed to be a direct or indirect measure of decoding (or word recognition), then read-aloud would clearly constitute a test modification. However, it is not clear if audio presentation changes the construct being measured when the construct is defined as comprehension rather than a combination of comprehension and decoding. Phillips (1994) argues that measurement specialists should consider the impact of modifications on the constructs measured and the validity of the resulting test scores. Assuming that an examinee with a disability is incapable of adapting to the standard testing administration, Phillips argues that any changes to testing conditions should be avoided if the change would (a) alter the skill being measured, (b) preclude the comparison of scores between examinees that received accommodations and those who did not, or (c) allow examinees without disabilities to benefit (if they were granted the same accommodation). This last criterion is debatable, and several researchers have argued that accommodations should be provided if they offer a differential boost to students with disabilities (Elliott & McKevitt, 2000; Fuchs & Fuchs, 1999; Pitoniak & Royer, 2001).

More recently, Sireci, Scarpati, and Li (2005) have termed the investigation of this differential performance as the *interaction hypothesis*. Both the interaction hypothesis and the differential boost argument indicate that an accommodation may still be considered valid if students with disabilities benefit differentially more than students without disabilities. This argument has been criticized as not focusing on the predictive validity of accommodated and nonaccommodated test scores and for the potential that ceiling effects can reduce the observed performance gains in the higher performing comparison group (Koenig & Bachman, 2004).

Several studies, however, have used the differential boost framework to examine the impact of audio presentation on tests of mathematics. (See Sireci, Scarpeti, and Li, 2005, and

Tindal & Fuchs, 2000, for complete reviews of the research.) All of these studies provided some evidence that audio presentation does result in a differential performance boost on tests of mathematics. In addition to the mathematics studies, five studies have examined the impact of audio presentation on tests of reading comprehension using a differential boost framework. A summary of this research is provided in the following section.

Review of Research

Kosciolek and Ysseldyke published the first differential boost study of a read-aloud accommodation on a test of reading in 2000. They used a quasi-experimental design to compare performance among third through fifth grade students with and without disabilities on a norm-referenced test of reading. Results indicated no significant difference in performance gains due to the read-aloud accommodation; however, the study faced several limitations, including a small sample size (n = 31) that limited the researchers' ability to detect significant differences.

In a second study, Meloy, Deville, and Frisbie (2002) examined the performance of middle school students without disabilities (NLD) and middle school students with RLDs. The sample size (n = 260) was larger than the Kosciolek and Ysseldyke study, but most students (76%) did not have a disability and students did not participate in both conditions (standard and audio). Instead students were randomly assigned to either audio or standard and took all content areas under the same condition. Results indicated similar performance gains for students with and without disabilities.

The third study (McKevitt & Elliott, 2003) had a small sample (n = 39) of eighth grade students that was split between students with and without disabilities. The sample of students with disabilities was limited to students who were receiving special education services in reading/language arts. All students took two reading assessments: one with no accommodations and one with teacher-recommended accommodations and audio presentation. The accommodated administration was done in a small group of students who received the same package of accommodations (e.g., extra time with audio presentation). Audio presentation was delivered via an audiocassette recording of the test read at a rate of 170 words per minute. The tape could be paused to allow students to record answers, but test content was not repeated. The researchers divided the TerraNova Multiple Assessments reading test into two forms (each form included 17 multiple choice items and either 2 or 4 constructed response items). Because equated test forms were not used, raw scores were converted to normal curve equivalent scores on a common scale.

A repeated measures analyses of variance (RM-ANOVA) was conducted to test for significant differences between students with and without disabilities on the two measures (with accommodations and without). Results indicated no significant performance differences. Small differences in effect sizes (0.22 to 0.25) were found for both the students with disabilities and the students without disabilities when comparing the difference between accommodated and nonaccommodated test scores.

A fourth study, by Crawford and Tindal (2004), examined the effects of read-aloud on a standardized test of reading for fourth and fifth grade students with and without learning disabilities. The sample size was large ($N = 338$), but most of the students (78%) did not have a disability. The audio presentation was a group administered video, and the students could not hear the passage or test questions repeated. The two 30-item forms included reading comprehension items that were assembled from a larger pool of items previously developed for a state assessment. Time limits were liberal (45 minutes for 30 questions) for both test sessions (with read-aloud and without read-aloud). An analysis of variance found no significant difference in performance by test form, order of accommodation administered (standard first or read-aloud first), or grade level (fourth or fifth). An RM-ANOVA revealed a disability by accommodation interaction, indicating a differential performance boost from the audio presentation accommodation compared to the standard administration for students with learning disabilities, relative to students without disabilities.

The fifth and most recently reviewed differential boost study of read-aloud examined the interaction hypothesis using a third grade state reading assessment (Fletcher et al., 2006). Nearly 200 students were randomly assigned to take a practice form of the Grade 3 Texas Assessment of Knowledge and Skills (TAKS) reading assessment under standard conditions or with accommodations. The accommodated condition consisted of a bundle of three accommodations (i.e., extending testing across two sessions, reading of proper nouns aloud, and reading the question stems and answer choices aloud). In addition, students were administered individual assessments of oral language vocabulary (i.e., Picture Vocabulary subtest from the Woodcock Language Proficiency Battery-Revised) and decoding (Letter-Word Identification and Word Attack subtests from the Woodcock-Johnson III Test of Achievement). The decoding measure was used to select students for participation in the study, so the sample of students with a disability (n = 91) only included poor decoders and the sample of students without a disability (n = 91) only included average decoders.

An analysis of variance was conducted to examine performance by level of decoding ability, accommodated condition (standard versus accommodated), and level of vocabulary knowledge. The first analysis examined the decoding ability by accommodation interaction with vocabulary score as a covariate and found a statistically significant interaction between decoding ability and accommodation. The authors concluded that poor decoders received a differential boost from the accommodated version when compared to average decoders. The effect size for this difference was large (d = 0.91) for poor decoders and small (d = 0.15) for average decoders. In addition, they noted that students with higher vocabulary scores had higher performance on the TAKS but that vocabulary score did not significantly interact with decoding ability or accommodation. A secondary analysis examined the performance differences within the poor decoding group using decoding score as a covariate and found no significant effect for the accommodation. The authors concluded that the severity of decoding difficulties within the poor decoding group was not related to the effects of the accommodations. While this study provided additional information on how decoding ability impacts differential boost and had a relatively large sample size, the study provided no information on the full range of students with and without RLDs (e.g., students with RLDs who are average decoders or students without disabilities who are poor decoders).

Limitations of Prior Research

Of the five studies reviewed, two found evidence of differential boost from a read-aloud accommodation and three did not find any evidence of differential boost. All five studies, however, have one of several limitations: the sample size was too small to detect significant differences, the study did not use a repeated measures design, or the subgroup of students with disabilities was poorly defined. In addition, none of the studies examined the validity of test scores taken with and without a read-aloud accommodation. While the small sample sizes and repeated measures design are relatively easy to remedy in future research, the final limitation (poorly defined disability subgroup) is more difficult to remedy without testing students on their decoding or fluency ability (as Fletcher et al., 2006, did).

Research Questions

Although previous differential boost studies have provided some information on the impact of audio presentation on test scores, the inconsistent findings and the limitations of several of the studies (poorly defined subgroups, small sample sizes, and lack of repeated measures design) suggest the need for additional research. In this study, we have a sufficient sample size to examine the interaction hypothesis, and we have also collected data to account for individual differences in reading fluency (a measure of reading speed and accuracy that is correlated with reading comprehension but is also a key indicator of the word-level and fluency-level reading disability subtypes described by Fletcher et al., 2006). This study uses a randomized-within-subject design (each student taking both standard and audio format tests) to examine (a) the interaction model for differential boost at two grade levels (fourth and eighth), (b) the influence of reading fluency ability and ceiling effects on those results, and (c) the validity of test scores using teachers' ratings of reading comprehension as an alternate measure of performance.

Method

Sample

Selection of Schools

A total of 84 schools participated (11 schools containing both fourth and eighth grade students, 45 schools containing only a fourth-grade group, and 28 schools containing only an eighth-grade group). Participating schools received score reports for each student and an honorarium. A total of 2,691 public and private schools in New Jersey were asked to participate. Of these, 11.2% indicated an interest in participating and 3.5% were included in the final sample. The final sample of schools was selected to represent socio-economic and ethnic diversity; however, preference was given to schools with larger numbers of students with learning disabilities.

Selection of Students

All fourth and eighth grade students with RLDs in participating schools were asked to participate in this study. The school coordinator was instructed to select those students who had been specifically identified in their Individualized Educational Plan (IEP) as having an RLD. In addition, school coordinators were asked to omit students with multiple disabilities (e.g., Attention

Deficit Hyperactive Disorder and learning disabilities). Of the students with RLDs selected for participation, 65% participated. Once the RLD sample was identified, a slightly larger number of students without a disability were randomly selected from an alphabetical list of students in the fourth or eighth grade at the same school. Of this sample, 65% participated in the study.

Description of Final Sample

The full sample for this study included 1,181 fourth grade students (527 with RLD and 654 with NLD) and 847 eighth grade students (376 with RLD and 471 with NLD). The racial/ethnic diversity was nearly identical across grades and disability categories; however, the percentage of Asian NLD students was larger (8.4%) than the percentage of Asian RLD students (3.6%). The racial/ethnic percentages by grade and disability category are displayed in Table 1. The sample of NLD students was evenly distributed by gender, but there were more boys in the sample of students with RLDs (66% in Grade 4 and 56% in Grade 8), which is consistent with national data.

Table 1

Percentage of Students by Race, Grade, and Disability Category

	Grade			Grade		
	4	8	Total	4	8	Total
Ethnicity		RLD			NLD	
White	63.7	59.7	62.1	62.1	58.1	60.5
Black	15.2	14.6	14.9	12.1	14.3	13.1
Hispanic	18.1	21.1	19.4	17.2	18.9	17.9
Asian	3.0	4.6	3.6	8.5	8.1	8.4
Other	0.0	0.0	0.0	0.0	0.4	0.2

Note. RLD = students with reading-based disability, NLD = students with no reading-based disability.

Teachers of RLD students were also asked to describe the aspect of reading that was impacted by the student's disability. The percentage distribution is reported in Table 2, with about half of the teachers responding that the students had problems with a combination of comprehension and decoding or word recognition and that a very large percentage of students had problems with comprehension (74% of RLD fourth graders and 83% of RLD eighth graders). This distribution is significant because this study is examining performance gains on a test of comprehension and the performance gains from audio presentation may be different for students with comprehension problems than for students with problems in decoding or word recognition (and no problems with comprehension).

Table 2

Number and Percent of Reading-Based Learning Disabilities by Aspect of Reading Impacted by Disability

Aspect of reading impacted by disability	Grade 4 RLD		Grade 8 RLD	
	n	%	*n*	%
Comprehension + decoding + other	43	8%	36	10%
Comprehension + decoding	256	49%	148	40%
Comprehension + other	7	1%	20	5%
Comprehension	84	16%	103	28%
Decoding	39	7%	17	5%
Decoding + other	4	1%	2	1%
Other	16	3%	8	2%
None/NR	74	14%	34	9%

Note. RLD = students with reading-based disability, NLD = students with no reading-based disability, NR = no response.

Materials

Research materials included two equated forms of the Gates-McGinitie Reading Tests (GMRT) Fourth Edition (Reading Comprehension subtest only), one form of the Woodcock-Johnson III Diagnostic Reading Battery (WJ-III DRB) Reading Fluency subtest (Woodcock, Mather, & Schrank, 2004), the one form of the Test of Silent Word Reading Fluency (TOSWRF), a student roster with demographic information, a student survey, and a teacher survey. In addition, the fourth grade sample was administered two additional subtests from the WJ-III DBR (Letter-Word Identification and Word Attack).

Assessments

Reading comprehension test. The GMRT Reading Comprehension subtest included two parallel equated forms (S and T) with short reading passages followed by multiple-choice reading comprehension questions for Grades 4 and 7/9. Each passage has three to six questions for a total of 48 questions per form. Since GMRT has a vertical scale across grades, the average scale scores for Grade 8 are higher than the average scale scores for Grade 4. In order to isolate any performance gain due to read-aloud, the standard administration included two accommodations commonly used by students with learning disabilities: time and a half extra time and answers recorded in the test booklet instead of on an answer sheet. The audio administration included time and a half extra time, answers recorded in the test booklet, and audio presentation. To increase consistency in the audio presentation, it was delivered using a compact disc (CD) player with headphones. The passage and each test question with answer choices were recorded on separate tracks and students were allowed to replay the tracks. Passages were read at a rate of 150 to 160 words per minute. Students had access to the test form in paper format as well as being able to listen to it.

Fluency assessments. Two measures of reading fluency were group administered to all students in the sample. These tests include the WJ-III DRB Reading Fluency subtest and the TOSWRF Form A. The WJ-Reading Fluency subtest requires the student to read simple sentences and mark the statement as true (yes) or false (no). They must complete as many items correctly as they can within a 3-minute time limit. The median reliability is 0.90 for ages 5 to 19 (Woodcock, Mather, & Shrank 2004). WJ-Reading Fluency raw scores were converted to W-scores for analyses. W-scores are calculated on an equal interval scale as an intermediary step and are recommended by the test publisher for use when conducting research studies (Shrank, Mather, & Woodcock, 2004, page 71). The TOSWRF is designed to measures student's ability to recognize printed English words and requires students to look at a stream of English words that are not separated by spaces (e.g., inatothe) and place slash lines between as many words as possible within 3 minutes (e.g., in/a/to/the). Raw scores were then converted to standardized scores based on norms from the test manual. The mean reliability is 0.92 for ages 7 to 17 (Mather, Hammill, Allen, & Roberts, 2004)

Word recognition assessments. Fourth grade students were also administered two additional subtests from the WJ-III DRB: Letter-Word Identification and Word Attack

(Woodcock, Mather, & Shrank, 2004). Together, these subtests make up the Basic Reading Skills Cluster, which the comprehensive manual describes as an aggregate measure of sight vocabulary, phonics, and structural analysis. The cluster has a median reliability of 0.92 among 5 to 19 year olds (0.91 and 0.87 for Letter-Word Identification and Word Attack tasks respectively; Shrank, Mather, & Woodcock, 2004). For the Word Attack subtest, individuals are asked to read aloud a set of letter combinations that are phonically consistent patterns of English orthography but are nonwords or low-frequency words. In the Letter-Word Identification task, individuals must pronounce correctly a set of English words. In both tasks, the items become increasingly more difficult across the list of items, and the task is terminated when the individual's responses are incorrect on a set number of items in a row. In addition, the examiner will return to easier items until a set number of items are answered correctly in a row. Raw scores for both subtests were then converted to W-scores for analyses.

Surveys

Teacher survey. The teacher survey included questions regarding the students' disability classification, classroom setting, and accommodations they typically receive on standardized tests and in the classroom. In addition, teachers rated each student's listening and reading comprehension ability relative to the students they teach as well as "a typical fourth (or eighth) grade student." Ratings of listening comprehension were collected to examine if test scores obtained from the audio presentation accommodation were more highly correlated with teachers' ratings of listening comprehension than with teachers' ratings of reading comprehension. Finally, teachers were asked to predict the test format (audio or standard) on which each student would perform better and to indicate which components of reading would be impacted by each student's RLD.

Student survey. The student survey included five short questions that were read aloud to students following completion of both test forms. Questions focused on what parts of the CD they listened to, their reading rate relative to the pace of the CD, if they liked to read, which format (audio or standard) they preferred, and which format they thought they did better on. Survey responses were used only to ensure that students included in the final sample reported that they had listened to the audio version of the test.

Student roster. The student roster was completed by the school coordinator and the data collection team leader and included demographic information (e.g., students' disability status, race, date of birth, and level of English language proficiency) and indicated experimental group assignments.

Procedure

Each participating school was assigned to one of two accommodation orders that varied in which accommodation (audio or standard) was presented first. Students were then randomly assigned to one of two form orders (Form S first versus Form T first). This resulted in four possible experimental groups that varied in the order in which the students received the two test forms and the order in which they received the accommodation conditions (see Table 3). Each NLD and RLD student at Grades 4 and 8 was assigned to one of the four experimental groups. In this within-subject design, the test form and accommodation condition were counter-balanced (to reduce the impact of any accommodation order or test form effect) and all students took two forms of the reading test (one with and one without an audio presentation accommodation). Extra time and recording answers in the test booklet were given under both conditions to ensure that neither confounded the interpretation of the results.

Table 3

Design for Gates-McGinitie Test Administration

Group	Session 1		Session 2		Group abbreviation
	Form	Accommodation	Form	Accommodation	
1	S	Standard	T	Audio	SSTA
2	S	Audio	T	Standard	SATS
3	T	Standard	S	Audio	TSSA
4	T	Audio	S	Standard	TASS

Data analyses included RM–ANOVA, comparing performance on the two measures (audio and standard) by group (RLD and NLD) and by test form/order of condition (based on the four experimental groups in Table 3). The sample size by experimental group, grade, and disability status are reported in Table 4. In addition to the RM-ANOVAs, a set of repeated measures analysis of covariance (RM-ANCOVA) with reading fluency as a covariate was conducted to examine the impact of poor reading fluency on the interaction hypothesis. The covariate used in the RM-ANCOVA was selected after examining the intercorrelations between the GMRT administered without read-aloud standard GMRT and all supplemental tests administered. Finally, a set of RM-ANOVAs were conducted to eliminate students who scored at the top of the distribution on the standard test. This last set of analyses was conducted to test for a potential ceiling effect in the NLD population that could possibly mask the differential performance gains in this population (a concern raised by Koenig and Bachman, 2004).

Table 4

Sample Size by Experimental Group, Grade, and Disability Status

Group	Grade 4		Grade 8	
	RLD	NLD	RLD	NLD
1	136	160	99	121
2	132	169	78	122
3	137	159	100	115
4	122	166	99	113

Note. RLD = students with reading-based learning disability,
NLD = students with no learning disability.

In addition to the RM-ANOVAs, this study used analysis of correlational data and regression procedures to examine the predictive validity of test scores (audio, standard, and fluency) relative to teachers' ratings of reading comprehension and listening comprehension by grade and disability status (RLD and NLD). Although limitations exist in the reliability and accuracy of teachers' ratings, these analyses provide some information on the validity of test scores, which is lacking in prior research on the impact of read-aloud accommodations. A final group of analyses examined the accuracy of teachers' predictions about which test format (audio or standard) would result in the best score for RLD and NLD students at each grade level.

Results

Differential Boost

We initially performed RM-ANOVAs that showed no significant interactions between disability status and either form order or accommodation order (see Appendix A for RM-ANOVA by disability status, form order, and accommodation order). Based on these results, we combined the test form order and accommodation order into one variable (experimental group). On average, the RLD group had lower test scores and a larger boost from the audio presentation. (See Table 5 for mean scores by grade and disability status.)

Table 5

Means and Standard Deviations for Scaled Scores Gates-McGinitie Reading Test (GMRT; Standard, Audio, and Boost) and Woodcock-Johnson III Diagnostic Reading Battery (WJ-III DRB) Reading Fluency Raw Score by Grade and Disability Status

	Grade 4			
	RLD (*n* = 527)		NLD (*n* = 654)	
	M	SD	M	SD
Standard	456.6	32.0	496.9	37.5
Audio	476.7	30.0	501.9	32.5
Boost	20.1	29.0	5.0	23.7
Fluency	473.3	20.7	500.4	24.6
	Grade 8			
	RLD (*n* = 376)		NLD (*n* = 471)	
Standard	510.8	27.6	552.8	32.9
Audio	520.6	27.3	554.7	30.5
Boost	9.8	22.9	2.0	20.8
Fluency	513.6	33.6	560.0	41.7

Note. RLD = students with reading-based learning disability, NLD = students with no learning disability.

The RM-ANOVAs (see Tables 6 and 7) indicated that the entire sample showed a significant performance boost on the audio version at Grade 4 (F [1, 1173] = 265.81, $p < .001$) and Grade 8 (F [1, 839] = 62.84, $p < .001$). In addition, a differential boost was also found at Grade 4 (F [1, 1173] = 96.46, $p < .001$) and Grade 8 (F [1, 839] = 27.88, $p < .001$) with RLD students having a larger boost than NLD students. Also in Grade 8, a significant interaction was noted between experimental group and boost (F [3, 839] = 11.87, $p < .001$), but no interaction was found among disability status, experimental group, and boost. The significant interaction of boost by experimental group appears to be due to a smaller boost found in Group 3 (TSSA; see Table 3) and a larger boost found for Group 4 (TASS; see Table 3) for both the NLD and RLD groups (see Appendix B for means and standard deviations by experimental group, grade, and disability status). In an attempt to explain this effect, we have looked for a school effect as well as for students with unexpected response patterns but found none. Further research will examine differential item functioning (DIF) across the groups.

Table 6

Repeated Measures Analysis of Variance (RM-ANOVA) for Grade 4

Source	*df*	*MS*	*F*	*p*
Within subjects				
Boost	1	91,134.66	265.81***	.000
Boost x RLD	1	33,070.35	96.46***	.000
Boost x experimental group	3	212.68	0.62	.602
Boost x RLD x experimental group	3	461.67	1.35	.258
Error	1,173	342.85		
Between subjects				
Boost x RLD	1	627,600.12	337.26***	.000
Boost x experimental group	3	5,483.82	2.95*	.032
Boost x RLD x experimental group	3	3,617.51	1.94	.121
Error	1,173	1,860.86		

Note. RLD = reading-based learning disability.

*p < .05. *** p < .001.

Because the RLD and NLD populations were not of equal ability, we also conducted RM-ANOVAs separately for each population (i.e., Grade 4 RLD, Grade 4 NLD, Grade 8 RLD, and Grade 8 NLD). Results of these analyses were very similar to those reported above with a significant but smaller boost found for the NLD sample than for the RLD sample. An experimental group by boost interaction was found at Grade 8 (for both RLD and NLD) but not Grade 4. Results of these RM-ANOVAs are reported in Appendix C.

Although a significant interaction between boost and disability status (boost x RLD in Tables 6 and 7) was found, the effect sizes of the boost were small to medium in size: 0.33 and 0.18, for all fourth and eighth grade students respectively. Because the boost was significantly different for the RLD and NLD samples, we also computed effect sizes for each population separately. Results showed the amount of boost based on disability status: 0.57 and 0.14 for RLD and NLD respectively at Grade 4 and 0.32 and 0.06 for RLD and NLD respectively at Grade 8. The standard deviation used to calculate the effect size was computed from the weighted pooled variances for the RLD and NLD samples on the standard (nonaudio) condition. Due to the difference in the score distributions for the RLD and NLD samples on the standard condition, the standard deviation (SD) when calculated directly with both samples was artificially high (higher than either groups' SD), so pooling the variances better represented the distribution of scores.

Table 7

Repeated Measures Analysis of Variance (RM-ANOVA) for Grade 8

Source	*df*	*MS*	*F*	*p*
Within subjects				
Boost	1	14,326.93	62.84***	.000
Boost x RLD	1	6,356.55	27.88***	.000
Boost x experimental group	3	2,707.13	11.87***	.000
Boost x RLD x experimental group	3	58.86	0.26	.856
Error	839	227.99		
Between subjects				
Boost x RLD	1	599,883.97	385.29***	.000
Boost x experimental group	3	746.96	0.48	.696
Boost x RLD x experimental group	3	737.82	0.47	.701
Error	839	1,556.96		

Note. RLD = reading-based learning disability.

*** $p < .001$.

Differential Boost Controlling for Fluency

Because students in the RLD group had lower reading fluency scores than most students in the NLD group, we conducted a RM-ANCOVA with reading fluency as the covariate. Reading fluency consisted of standardized W-scores from the WJ-Reading Fluency subtest. The WJ-Reading Fluency measure was selected as the covariate because it showed the highest correlation with standard score for all four of the subgroups (RLD Grade 4, RLD Grade 8, NLD Grade 4, and NLD Grade 8). The other reading subtests we administered (WJ-Word Attack, WJ-Letter-Word

Identification, and the TOSWRF) had lower correlations with the standard scores. The complete set of correlation tables for all subtests administered by group are included in Appendix D.

The RM-ANCOVAs are reported in Tables 8 and 9. Results showed a significant differential boost by RLD when controlling for fluency at both Grade 4 (F [1, 1173] = 22.50, $p < .001$) and Grade 8 (F [1, 831] = 11.08, $p < .001$), although the boost effect is somewhat reduced when compared to the previous analyses which did not control for fluency.

Table 8

Repeated Measures Analysis of Covariance (RM-ANCOVA) for Grade 4 With Fluency as a Covariate

Source	*df*	*MS*	*F*	*p*
Within subjects				
Boost	1	23,072.88	71.43***	.000
Boost x fluency (covariate)	1	19,017.42	58.87***	.000
Boost x RLD	1	7,269.45	22.50***	.000
Boost x experimental group	3	292.34	0.91	.438
Boost x RLD x experimental group	3	484.59	1.50	.213
Error	1,171	323.03		
Between subjects				
Fluency (covariate)	1	746,236.29	610.56***	.000
Boost x RLD	1	59,878.20	48.99***	.000
Boost x experimental group	3	2,507.122	2.05	.105
Boost x RLD x experimental group	3	1,029.24	0.84	.471
Error	1,171	1,222.22		

Note. *RLD = reading-based learning disability.

** $p < .001$.

Differential Boost Controlling for Ceiling Effect

Because the distribution of test scores for the NLD groups are skewed toward the top of the distribution, a ceiling effect could be reducing the observed boost for this sample. For this reason, we repeated the RM-ANOVA after eliminating students who answered more than 45 items correct on the standard administration. This included 36 fourth graders (5 RLD and 31 NLD) and 29 eighth graders (all NLD). Results were nearly identical to those reported for the full sample, indicating that possible ceiling effects do not appear to influence results. Detailed results are provided in Appendix E.

Table 9

Repeated Measures Analysis of Covariance (RM-ANCOVA) for Grade 8 With Fluency as a Covariate

Source	*df*	*MS*	*F*	*p*
Within subjects				
Boost	1	1,798.06	7.89**	.005
Boost x fluency (covariate)	1	1,152.15	5.06*	.025
Boost x RLD	1	2,525.47	11.08**	.001
Boost x experimental group	3	2,531.70	11.11***	.000
Boost x RLD x experimental group	3	60.23	0.26	.851
Error	831	227.90		
Between subjects				
Fluency (covariate)	1	306,553.13	257.58***	.000
Boost x RLD	1	140,726.57	118.25***	.000
Boost x experimental group	3	1,321.00	1.11	.344
Boost x RLD x experimental group	3	671.36	0.56	.639
Error	831	1,190.11		

Note. RLD = reading-based learning disability.

*$p < .05$, **$p < .01$, *** $p < .001$.

Predictive Validity of Audio and Standard Scores

To examine the predictive validity of both the standard and audio test scores, we conducted two sets of analyses: one based on correlational analyses and the other based on regression analyses. The correlational analyses compared the correlation of test scores (standard and audio) with teachers' ratings of comprehension. Both sets of analyses (correlational and regression) are limited by the reliability and validity of teachers' ratings of reading comprehension and how teachers define reading comprehension. These analyses do, however, provide some insight into the predictive validity of test scores taken with and without read-aloud accommodations that are missing from prior research. Both sets of analyses are included in this report.

Correlations

In the case of the correlational analyses, we first examined if the audio and standard GMRT scores were more highly correlated in the NLD population than in the RLD population. We hypothesized that if the audio and standard scores were highly correlated in the NLD population, then these scores could possibly be measuring a single construct of comprehension of text for NLD test-takers. Likewise, we hypothesized that if the audio and standard scores were not so highly correlated in the RLD population, then these scores could possibly be measuring different constructs or a combination of constructs (comprehension of text, decoding, word recognition, and reading fluency) for RLD test-takers.

The next set of correlations attempted to determine which test score (standard GMRT or audio GMRT) was more highly correlated with teachers' ratings of reading comprehension for each subgroup. In addition, we examined if the audio score was more highly correlated with teachers' ratings of listening comprehension than teachers' ratings of reading comprehension, which would give support to the assertion by some states (e.g., Wisconsin) that a read-aloud accommodation changes the test from a test of reading comprehension to a test of listening comprehension.

Correlations between test scores. Tables 10 and 11 display the Pearson correlation coefficients among test scores, boost, and teachers' ratings of reading and listening comprehension by grade and disability status. For both grades, the correlation between test scores derived under standard conditions (standard GMRT) and test scores derived under the accommodated audio condition (audio GMRT) is higher for the NLD group than the RLD group (0.78 compared to 0.56 at Grade 4 and 0.79 compared to 0.65 at Grade 8). In addition, the correlations for the NLD group between audio and standard are similar to the correlations between forms as reported by the tests' technical manual (0.86 for Forms S and T at Grade 4 and 0.84 for eighth graders on Forms S and T at Grade 7/9 (see MacGinitie, MacGinitie, Maria, & Dreyer, 2000b]. These relationships indicate that the audio and standard administration may be measuring a similar construct(s) for the NLD population but perhaps a somewhat different construct(s) for the RLD population, particularly at Grade 4.

Table 10

Intercorrelations Between Test Scores, Boost, and Teachers' Ratings for Grade 4

Measure	1	2	3	4	5	6	7	8
RLD (*N* = 508)								
1. Standard	--							
2. Audio	.56	--						
3. Boost	-.52	.41	--					
4. Fluency	.58	.38	-.25	--				
5. TR-reading other	.31	.19	-.15	.34	--			
6. TR-reading typical	.46	.37	-.13	.52	.72	--		
7. TR-listening other	.12	.16	.03	.14	.57	.37	--	
8. TR-listening typical	.30	.33	.02	.33	.38	.52	.76	--
NLD (*N* = 635)								
1. Standard	--							
2. Audio	.78	--						
3. Boost	-.51	.14	--					
4. Fluency	.60	.55	-.20	--				
5. TR-reading other	.58	.57	-.15	.52	--			
6. TR-reading typical	.61	.58	-.17	.54	.91	--		
7. TR-listening other	.45	.44	-.10	.38	.74	.71	--	
8. TR-listening typical	.45	.44	-.11	.39	.73	.76	.94	--

Note. The sample size is slightly reduced due to incomplete teacher survey data for some students in the sample. TR = teachers' ratings; RLD = reading-based learning disability; NLD = no learning disability.

Table 11

Intercorrelations Between Test Scores, Boost, and Teachers' Ratings for Grade 8

Measure	1	2	3	4	5	6	7	8
				RLD (*N* = 368)				
1. Standard	--							
2. Audio	.65	--						
3. Boost	-.43	.40	--					
4. Fluency	.47	.33	-.17	--				
5. TR-reading other	.32	.24	-.09	.19	--			
6. TR-reading typical	.41	.30	-.12	.34	.63	--		
7. TR-listening other	.26	.24	-.01	.13	.70	.39	--	
8. TR-listening typical	.39	.30	-.11	.31	.48	.68	.65	--
				NLD (*N* = 458)				
1. Standard	--							
2. Audio	.79	--						
3. Boost	-.43	.22	--					
4. Fluency	.48	.49	-.03	--				
5. TR-reading other	.51	.54	-.02	.38	--			
6. TR-reading typical	.52	.54	-.03	.41	.91	--		
7. TR-listening other	.44	.47	-.01	.36	.83	.80	--	
8. TR-listening typical	.47	.49	-.04	.37	.80	.85	.94	--

Note. The sample size is slightly reduced due to incomplete teacher survey data for some students in the sample. TR = teachers' ratings, RLD = reading-based learning disability, NLD = no learning disability.

Correlations between teachers' ratings and test scores. In addition, we examined the relationships between teachers' ratings of both listening comprehension and reading comprehension and test scores. Teachers were asked to rate their students' reading and listening comprehension compared to "a typical fourth (or eighth) grader" and "your other students." Ratings were on a 5-point Likert Scale, which included Significantly Below Average, Below Average, Average, Above Average, and Significantly Above Average. Our purpose for asking questions two different ways (i.e., compared to a typical student and compared to other students) was to reduce the impact of school and classroom ability effects on teacher ratings. It appears that this did have some success, since the teachers' ratings of student performance compared to a typical fourth (or eighth) grader were more highly correlated with the GMRT scores than teachers' ratings of performance compared to their other students. In addition, the correlations between teachers' other students and a typical student are very high (0.90 and higher) for the NLD sample and somewhat lower (0.60 to 0.70) for the RLD sample. Based on this information, we used teachers' ratings of comprehension compared to a typical student to examine the correlations between test scores and teachers' ratings, and in the regression analyses that follow.

Results (displayed in Tables 10 and 11) indicate that the correlations between teachers' ratings and test scores (standard and audio) are slightly higher for the NLD sample than for the RLD sample for both grades. For example, the correlations between test scores and teachers' ratings range from 0.44 to 0.61 for Grade 4 NLD compared to a range of 0.30 to 0.46 for Grade 4 RLD. In addition, the correlation between teachers' ratings of reading comprehension and test scores under the standard condition is higher than the correlation between teachers' ratings of listening comprehension and test scores from the audio condition for three groups (Grade 4 RLD, Grade 4 NLD and Grade 8 RLD) and similar for Grade 8 NLD. This lower correlation may have a variety of causes that stem from a teacher's definition of listening comprehension (i.e., listening to a class lecture rather than listening to reading passages read aloud) or from the fact that the audio score was intertwined with reading comprehension because students were provided with a print copy of the text that was read aloud via the CD player. The low correlation between audio score and teachers' ratings of listening comprehension, however, does not support the argument that test scores obtained under the read-aloud accommodation are a better measure of listening comprehension than a measure of reading comprehension or that the test is measuring listening comprehension (as teachers define the construct) instead of reading comprehension as some states have asserted.

Regression Analyses

Based on the results of the correlational analyses, this next set of analyses (regression) used teachers' ratings of reading comprehension compared to a typical fourth (or eighth) grader as the external criterion measure of the construct being assessed. The external criterion we used (teachers' ratings) is not ideal due to the lack of accuracy in teacher predictions observed as well as the variations among teachers (e.g., special education teachers and the regular education teachers may vary systematically in their ability to predict student performance). However, we feel that it is important to conduct these analyses to at least provide some preliminary information on the predictive validity of scores obtained under both audio and standard testing conditions.

The primary purpose of these regression analyses was to determine if measuring comprehension and fluency skills in isolation (i.e., reading comprehension assessment with audio accommodation and direct measure of reading fluency) resulted in a better measurement of reading comprehension than measuring these skills together (i.e., standard reading comprehension assessment). This set of analyses was directed at the argument made by many states that fundamental reading skills (e.g., decoding, word recognition, and reading fluency) are measured indirectly by the states' reading comprehension assessment; therefore, scores will not be counted for NCLB accountability purposes if read-aloud is used because it interferes with a construct being assessed. To investigate this claim, we used regression analyses to examine the amount of variance in teachers' ratings of reading comprehension that was captured by the standard GMRT (Model 1), the standard GMRT and WJ-Reading Fluency (Model 2), the audio GMRT (Model 3), and the audio GMRT with WJ-Reading Fluency (Model 4). Differences in the variance captured by each of the alternative models (Models 2, 3, and 4) relative to the model currently used by most states (Model 1) by disability and grade subgroups are summarized in Table 12; full analyses are displayed in Appendix F.

The primary purpose for examining both Models 1 and 3 was to compare the amount of variance in teachers' ratings captured by the current testing policy in many states. Model 1 captures the policy of states that do not permit read aloud accommodations (i.e., all students should be tested without audio presentation because decoding and/or fluency are standards that are measured indirectly on the state reading assessment), while Model 3 captures the policy of states that do allow read-aloud accommodations on tests of reading (i.e., allowing some students with disabilities to be tested with audio accommodations captures comprehension proficiency in isolation of decoding and fluency). Because many state standards include reading fluency and because reading fluency measures are relatively short to administer, we also included a direct measure of reading fluency in addition to comprehension in Models 2 and 4. Model 4 examines the variance that could be captured by assessing reading fluency and comprehension in isolation on state assessments, while Model 2 captures the variance from reading fluency (in isolation) as well as reading fluency and comprehension in combination.

Table 12

Comparison of Alternate Measurement Models to Model 1—Standard Gates-McGinitie Reading Test (GMRT)

	Model 1	Model 2		Model 3		Model 4	
Group	R^2	R^2	Difference in R^2	R^2	Difference in R^2	R^2	Difference in R^2
Grade 4 NLD	.368	.414	.045	.331	-.037	.400	.032
Grade 8 NLD	.276	.310	.034	.294	.018	.322	.046
Grade 4 RLD	.211	.310	.099	.136	-.075	.307	.096
Grade 8 RLD	.164	.195	.031	.088	-.076	.156	-.008

Note. Model 1 = standard; Model 2 = standard + fluency; Model 3 = audio; Model 4 = audio + fluency; Difference in R^2 = Model x - Model 1.

We hypothesized that for NLD students at Grades 4 and 8, Model 1 would capture amounts of variance (in teachers' ratings of reading comprehension) equal to each of the other models. This hypothesis was based on an assumption that the reading fluency levels for the majority of NLD test-takers would be sufficient enough for the assessment to capture variance in comprehension achievement, rather than the combined variance of fluency and comprehension. For RLD students, we hypothesized that both the audio GMRT and WJ-Reading Fluency scores (Model 4) and the audio GMRT alone (Model 3) would capture more variance than the standard GMRT (Model 1). This hypothesis was based on an indication from the differential boost analyses that the reading fluency levels for the majority of RLD test-takers were not sufficient enough for the standard GMRT to capture a student's true comprehension ability, hence the large boost in test scores on the audio GMRT. Since both Models 3 and 4 included a measure of comprehension in isolation, we hypothesized that these two models would capture more variance in teachers' ratings of reading comprehension than the standard GMRT (Model 1).

In order to investigate these different models, we completed two sets of regression analyses. The first set of analyses examined the variance captured by the standard administration of the GMRT (Model 1), followed by the variance captured by the standard GMRT and the WJ-Reading Fluency subtest (Model 2) for all four subgroups (RLD Grade 4, NDL Grade 4, RLD Grade 8, and NLD Grade 8). The second set of regression analyses compared the variance captured by the audio administration of the GMRT (Model 3), followed by the variance captured by the audio GMRT and the WJ-Reading Fluency subtest (Model 4) for the same four subgroups. All eight analyses are displayed in Appendix F.

Results of regression analyses. We completed the two sets of analyses described above for four subgroups (RLD Grade 4, NLD Grade 4, RLD Grade 8, and NLD Grade 8); results for all eight regression analyses are displayed in Appendix F. To test our hypotheses, we compared the difference in variance (R^2) captured by each of the four models by subgroups (see Table 12 for summary). Our hypothesis was that Model 3 (audio GMRT) and Model 4 (WJ-Reading Fluency and audio GMRT scores) would capture more variance than Model 1 (standard score) for the RLD population but equal variance for the NLD population, which would support assessing comprehension in isolation (i.e., with a read-aloud accommodation) and fluency in isolation for the RLD population. Table 12 summarizes the differences in the amount of variance (R^2) captured by Model 1 (GMRT standard) and the other three models.

The findings were fairly consistent with our first hypothesis that the standard score (Model 1) is an adequate measure of reading comprehension for NLD students at Grades 4 and 8. The difference in the amount of variance captured by Model 1 and the other three models was small (ranging from -0.037 to 0.046) for NLD students at both grades.

Results from comparisons between Model 1 and Model 3 (which only compared standard to audio) for RLD students (at both grades) did not support our hypothesis that audio scores would capture more variance in teachers' ratings of reading comprehension than standard scores. Instead, the results indicated that the audio score alone captured less variance in teachers' ratings of reading comprehension than the standard score alone, which is consistent with the results from the correlations between test scores and teachers' ratings described earlier.

In addition, examination of the variance captured by standard score and reading fluency (Model 2) and the variance captured by the audio score and reading fluency (Model 4) show that both these models captured more variance in teachers' ratings of reading comprehension than standard score alone (Model 1) for Grade 4 RLD students but not Grade 8 RLD students. These findings could support the direct measurement of reading fluency for Grade 4 RLD students, particularly when read-aloud accommodations are used on the state reading assessment. However, to our knowledge, no states are assessing fluency (in isolation) on a standards-based accountability assessment.

Conclusions

The results of this study support the argument that students with learning disabilities benefit differentially from read-aloud accommodations at both fourth and eighth grades even when reading fluency ability and ceiling effects are taken into account. The differential performance boost is greater in Grade 4 than Grade 8, which appears to be related to a decrease in the boost from audio presentation for both students with and without RLDs at the higher grade level. This decrease is consistent with prior research indicating that as word recognition becomes more fluent and automatized, listening comprehension becomes a stronger predictor of reading ability, though word recognition continues to contribute significant variance even in skilled readers (Carver 2003; Carver & David, 2001; Gough & Walsh, 1991).

Although students with RLDs do benefit differentially from audio presentation, the validity and interpretation of audio test scores is questionable. The prior research on the impact of the read-aloud accommodation did not attempt to examine the validity of test scores obtained with read-aloud. Although this study attempted to examine the validity of test scores relative to teachers' ratings of reading comprehension, results should be interpreted with caution because our external criterion (teachers' ratings of reading comprehension) had two limitations. First, teachers' ratings were collected early in the school year (October and November), so these ratings may not be as accurate as ratings collected later in the year. Second, the two populations had different ability distributions, so the ratings of the RLD students were skewed toward the lower end of the scale and resulted in a 3-point distribution (Significantly Below Average, Below Average, and Average) for 97% of the sample.

Even with these limitations, the information provides some useful and interpretable results. Results indicate the correlation between the standard score and teachers' ratings of reading comprehension are higher than the correlation of audio score and the same teachers' ratings. This finding suggests that the standard score may be a better measure of reading comprehension as it is defined by teachers. In addition, these analyses demonstrated that standard score is more highly correlated with teachers' ratings of reading comprehension than listening comprehension, which suggests that the audio accommodation does not change the assessment to a test of listening comprehension (as it is defined by teachers). The regression analyses indicated that measuring comprehension and reading fluency in isolation may result in a more valid test score (than using only the standard administration) for students with RLDs in fourth grade. In addition, these analyses indicated that standard score is an adequate measure of reading comprehension for NLD students at both Grades 4 and 8, but standard score alone captures less variance in teachers' ratings for RLD students at both Grades 4 and 8. Finally these analyses indicate that the audio score alone decreases the validity of test scores for RLD students at both Grades 4 and 8 when teachers' ratings of reading comprehension are the external criterion. Due to the limitations of the external validity criterion (teachers' ratings), this finding should be investigated in future research studies. Based on these findings, however, it may be advisable for states to consider adding a measure of reading fluency to tests of reading comprehension that are read aloud.

Limitations

There were a few limitations of this study that should be noted. The primary limitation, which was noted earlier, is the use of teachers' ratings as a criterion measure of performance when examining the validity of test scores (both audio and standard). Another limitation is that the reading comprehension assessment used in this study may not be generalizable to state reading assessments because the passages were relatively short and none of the passages required students to compare and contrast different reading passages. Another limitation is that students were assigned to testing condition (audio first or standard first) at the school level, so some school effects may be present, although none were noted during data analysis. Finally, the experimental group effect noted in Grade 8 indicates that the test forms interacted with the order of administration and format (audio or standard) in some way that is not easily explained.

Future Research

This study has provided a rich source of data to examine the factors that determine when an audio presentation accommodation is most beneficial and how listening and reading comprehension are related for students with and without RLDs. Future data analyses of this data will include (a) an examination of factors that contribute to score boost (e.g., standard score, decoding, fluency, classroom accommodations, teacher predictions, student preferences), (b) the relationship between listening and reading comprehension by grade and disability status, and (c) DIF across populations. While this study takes a first step in examining the validity of accommodated and nonaccommodated test scores, future research could expand on this study by collecting more accurate measures of reading to use as an external criterion in the validity analyses.

References

Americans With Disabilities Act of 1990, 42 U.S.C. § 12101 *et seq.* (1990).

Bolt, S., & Bielinski, J. (2003, April). *Isolating math item characteristics that differentially affect students with reading disabilities*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (Technical Rep. No. 31). Minneapolis, MN: National Center on Educational Outcomes.

Cahalan-Laitusis, C., Cook, L., & Aicher, C. (2003, April). *Examining test items for students with disabilities by testing accommodation*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Carver, R. P. (2003). The highly lawful relationship among pseudoword decoding, word identification, spelling, listening, and reading. *Scientific Studies of Reading, 7*(2), 105–126.

Carver, R. P., & David, A. H. (2001). Investigating reading achievement using a causal model. *Scientific Studies of Reading, 5*(2), 107–140.

Clapper, A. T., Morse, A. B., Lazarus, S. S., Thompson, S. J., & Thurlow, M. L. (2005). *2003 state policies on assessment participation and accommodations for students with disabilities* (Synthesis Rep. No. 56). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Crawford, L., & Tindal, G. (2004). Effects of a read-aloud modification on a standardized reading test. *Exceptionality, 12*(2), 89–106.

ETS. (2007). *California Standards Tests (CST) technical report Spring 2006 administration* (pp. 158). Retrieved April 3, 2007, fromwww.cde.ca.gov/ta/tg/sr/documents/startechrpt06.pdf

Elliott, S. N., & McKevitt B. (2000, April). *Testing accommodations decisions: Legal and technical issues challenging educators*. Paper presented at the annual conference of the American Educational Research Association, New Orleans, LA.

Fletcher, J. M., Francis, D. J., Boudousquie, A., Copeland, K., Young, V., Kalinowski, S., & Vaughn, S. (2006). Effects of accommodations on high stakes testing for students with reading disabilities. *Exceptional Children, 72*(2), 136–150.

Fuchs, L. S., & Fuchs, D. (1999). Fair and unfair testing accommodations. *School Administrator, 56*, 24–29.

Gough, P. B., & Walsh, M. A. (1991). Chinese, Phoenicians, and the orthographic cipher of English. In S. A. Brady & D. P. Shankweiler (Eds.), *Phonological processes in literacy: A tribute to Isabelle Y. Liberman* (pp. 199–210). Hillsdale, NJ: Lawrence Erlbaum Associates.

Individuals with Disabilities Education Act (IDEA) of 1997, 20 U.S.C. § 1412(a) (17) (A). (1997).

Individuals with Disabilities Education Act (IDEA), 20 U.S.C. § 1400 et seq. (2004).

Kosciolek, S., & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test* (Technical Rep. No. 28). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Koenig, J. A., & Bachman, L .F. (2004). Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assess. Washington, DC: National Academies Press.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000a). *Gates-MacGinitie Reading Tests.* Chicago, IL: Riverside.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000b). *Gates-MacGinitie Reading Tests, Fourth Edition* (Technical Rep. for Forms S & T.) Itasca, IL: Riverside Publishing.

Mather, N., Hammill, D. D., Allen, E. A., & Roberts, R. (2004). *TOSWRF: Test of Silent Word Reading Fluency: Examiner's manual.* Austin, TX: Pro-Ed.

McKevitt, B. C., & Elliott, S. N. (2003). Effects of perceived consequences of using read-aloud and teacher-recommended testing accommodation on a reading achievement test. *School Psychology Review, 23(*4), 583–600.

Meloy, L. L., Deville, C., & Frisbie, D. A. (2002). The effect of a read-aloud accommodation on test scores of students with and without a learning disability in reading. *Remedial and Special Education, 23*(4), 248–255.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 *et seq.* (2002).

Phillips, S. E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7*, 93–120.

Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research, 71*(1), 53–104.

Shrank, F. A., Mather, N., & Woodcock, R. W. (2004). *Comprehensive manual: Woodcock-Johnson III Diagnostic Reading Battery.* Itasca, IL: Riverside Publishing.

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*(4), 457–490.

Sireci, S., Li, S., & Scarpati, S. (2003). *The effects of test accommodation on test performance: A review of the literature* (Center for Educational Assessment Research (Rep. No. 485.) Amherst, MA: Center for Educational Assessment.

Thurlow, M., & Wiener, D. (2000). *Non-approved accommodations: Recommendations for use and reporting* (Policy Directions No. 11). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tindal, G., & Fuchs, L. (2000). *A summary of research on test changes: An empirical basis for defining accommodations*. Lexington: University of Kentucky, Mid-South Regional Resource Center Interdisciplinary Human Development Institute..

Wisconsin Department of Public Instruction. (2003). *Frequently asked questions about the wisconsin knowledge and concepts examinations at Grades 4, 8, and 10*. Retrieved February 7, 2006, from http://www.dpi.state.wi.us/oea/doc/wkce_faq10-04.doc

Woodcock, R., Mather, N., & Schrank, F. A. (2004). *Woodcock-Johnson III: Diagnostic Reading Battery.* Itasca, IL: Riverside Publishing.

List of Appendixes

Page

Appendix A - RM-ANOVA by Disability Status, Form Order, and Accommodation Order... 32

Appendix B - Means and Standard Deviations by Experimental Group, Grade, and Disability Status 37

Appendix C - Repeated Measures Analysis of Variance for Each Population......................... 39

Appendix D - Correlation Tables for All Subtests Administered by Group............................. 41

Appendix E - RM-ANOVA Results After Eliminating Top Performers................................. 43

Appendix F - Full Analyses of Variance in Alternative Models... 45

Appendix A

RM-ANOVA by Disability Status, Form Order, and Accommodation Order

The tables included in Appendix A are similar to Tables 6–9 included in the body of this report. The only difference is that one variable (experimental group) is divided into two variables (form order and accommodation order). For both grades, there was no significant interaction between boost and form order (see Table A1). At Grade 8, there was a significant interaction between boost and accommodation order (and boost x accommodation order x form order), but this difference did not interact with RLD classification (see Table A2). These findings were consistent after controlling for fluency (see Tables A3 and A4).

Table A1

Repeated Measures Analysis of Variance for Grade 4

Source	*df*	*MS*	*F*	*p*
Within subjects				
Boost	1	91,134.66	265.81***	.000
Boost x RLD	1	33,070.35	96.46***	.000
Boost x form order	1	391.82	1.14	.258
Boost x accommodation order	1	237.85	0.69	.405
Boost x RLD x form order	1	70.63	0.21	.650
Boost x RLD x accommodation order	1	1,004.27	2.93	.087
Boost x form order x accommodation order	1	0.00	0.00	1.000
Boost x RLD x form order x accommodation order	1	291.28	0.85	.357
Error	1,173	342.85		
Between subjects				
Boost x RLD	1	627,600.12	337.26***	.000
Boost x form order	1	1,374.26	0.74	.390
Boost x accommodation order	1	15,168.09	8.15**	.004
Boost x RLD x form order	1	6,515.92	3.50	.062
Boost x RLD x accommodation order	1	82.68	0.04	.833
Boost x form order x accommodation order	1	12.29	0.01	.935
Boost x RLD x form order x accommodation order	1	4,386.05	2.36	.125
Error	1,173	1,860.86		

Note.. RLD = reading-based learning disability.

** $p < .01$. *** $p < .001$.

Table A2

Repeated Measures Analysis of Variance for Grade 8

Source	*df*	*MS*	*F*	*P*
Within subjects				
Boost	1	14,326.93	62.84***	.000
Boost x RLD	1	6,356.55	27.88***	.000
Boost x form order	1	7.75	0.03	.854
Boost x accommodation order	1	4,089.21	17.94***	.000
Boost x RLD x form order	1	138.03	0.61	.437
Boost x RLD x accommodation order	1	7.56	0.03	.856
Boost x form order x accommodation order	1	3,839.50	16.84***	.000
Boost x RLD x form order x accommodation order	1	32.97	0.15	.704
Error	839	227.99		
Between subjects				
Boost x RLD	1	599,883.97	385.29***	.000
Boost x form order	1	126.43	0.08	.776
Boost x accommodation order	1	209.53	0.14	.714
Boost x RLD x form order	1	1,325.70	0.85	.356
Boost x RLD x accommodation order	1	741.63	0.48	.490
Boost x form order x accommodation order	1	1,881.57	1.21	.272
Boost x RLD x form order x accommodation order	1	264.62	0.17	.680
Error	839	1,556.96		

Note. RLD = reading-based learning disability.

*** $p < .001$.

Table A3

Repeated Measures Analysis of Covariance for Grade 4 With Fluency as a Covariate

Source	*df*	*MS*	*F*	*p*
Within subjects				
Boost	1	23,072.88	71.43***	.000
Boost x fluency (covariate)	1	19,017.42	58.87***	.000
Boost x RLD	1	7,269.45	22.50***	.000
Boost x form order	1	731.08	2.26	.133
Boost x accommodation order	1	130.59	0.40	.525
Boost x RLD x form order	1	195.54	0.61	.437
Boost x RLD x accommodation order	1	727.07	2.25	.134
Boost x form order x accommodation order	1	6.10	0.02	.891
Boost x RLD x form order x accommodation order	1	509.29	1.58	.209
Error	1,171	323.03		
Between subjects				
Fluency (covariate)	1	746,236.29	610.56***	.000
Boost x RLD	1	59,878.20	48.99***	.000
Boost x form order	1	8.39	0.01	.934
Boost x accommodation order	1	7,518.23	6.15*	.013
Boost x RLD x form order	1	1,163.52	0.95	.329
Boost x RLD x accommodation order	1	80.83	0.07	.797
Boost x form order x accommodation order	1	0.39	0.00	.986
Boost x RLD x form order x accommodation order	1	1,851.63	1.51	.219
Error	1,171	1,222.22		

Note. RLD = reading-based learning disability.

*** $p < .001$.

Table A4

Repeated Measures Analysis of Covariance for Grade 8 With Fluency as a Covariate

Source	*df*	*MS*	*F*	*p*
Within subjects				
Boost	1	1,798.06	7.89**	.005
Boost x fluency (covariate)	1	1,152.15	5.06*	.025
Boost x RLD	1	2,525.47	11.08**	.001
Boost x form order	1	20.22	0.09	.766
Boost x accommodation order	1	3,832.08	16.81***	.000
Boost x RLD x form order	1	154.52	0.68	.411
Boost x RLD x accommodation order	1	6.39	0.03	.867
Boost x form order x accommodation order	1	3,585.30	15.73***	.000
Boost x RLD x form order x accommodation order	1	22.22	0.10	.755
Error	831	227.90		
Between subjects				
Fluency (covariate)	1	306,553.13	257.58***	.000
Boost x RLD	1	140,726.57	118.25***	.000
Boost x form order	1	9.28	0.01	.930
Boost x accommodation order	1	349.61	0.29	.588
Boost x RLD x form order	1	1,171.50	0.98	.321
Boost x RLD x accommodation order	1	637.82	0.54	.464
Boost x form order x accommodation order	1	3,567.27	3.00	.084
Boost x RLD x form order x accommodation order	1	193.16	0.16	.687
Error	831	1,190.11		

Note. RLD = reading-based learning disability.

* $p < .05$. ** $p < .01$. *** $p < .001$.

Appendix B

Means and Standard Deviations by Experimental Group, Grade, and Disability Status

Tables B1 and B2 include the means, standard deviation, and sample size for students who completed all three reading measures (standard GMRT, audio GMRT, and WJ-Reading Fluency subtest) and the average performance boost (audio-standard) disability classification and experimental group for Grades 4 and 8 respectively. Results are consistent between experimental groups at Grade 4. However, in the Grade 8 sample, two of the experimental groups (3 and 4) are divergent in the degree of boost from audio for both the NLD and RLD groups.

Table B1

Means and Standard Deviations for Standard, Audio, Boost, and Woodcock-Johnson III Diagnostic Reading Battery (WJ-III DRB) Reading Fluency by Experimental Group, and Disability Status for Grade 4

			Standard		Audio		Boost		Fluency	
Disability/group		*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
NLD										
SSTA	1	160	501.6	36.2	505.0	28.9	3.5	25.2	43.0	12.6
SATS	2	169	496.0	38.2	504.8	34.9	8.8	24.6	42.2	11.9
TSSA	3	159	499.7	36.9	502.3	32.8	2.6	23.6	41.6	12.1
TASS	4	166	490.4	37.8	495.5	32.4	5.0	20.9	39.7	11.0
RLD										
SSTA	1	136	458.6	32.9	480.5	27.7	21.9	29.1	27.8	11.3
SATS	2	132	452.3	32.1	471.5	31.9	19.1	31.0	26.1	13.2
TSSA	3	137	458.7	31.2	478.2	28.7	19.5	30.7	27.4	10.2
TASS	4	122	456.8	31.5	476.4	31.1	19.6	24.6	27.5	11.0

Note. SSTA, SATS, TSSA, and TASS describe the test and accommodation order (see Table 3). NLD = no learning disability, RLD = reading-based learning disability.

Table B2

Means and Standard Deviations for Standard, Audio, Boost, and Woodcock-Johnson III Diagnostic Reading Battery (WJ-III DRB) Reading Fluency by Experimental Group, and Disability Status for Grade 8

			Standard		Audio		Boost		Fluency	
Disability/group		*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
NLD										
SSTA	1	121	554.4	29.2	556.9	32.0	2.4	20.5	68.5	15.9
SATS	2	122	549.2	33.1	552.1	28.2	2.9	18.0	65.9	16.5
TSSA	3	115	556.2	33.1	551.7	30.1	-4.5	19.3	67.4	14.8
TASS	4	113	551.3	36.1	558.4	31.2	7.0	23.8	67.3	17.4
RLD										
SSTA	1	99	512.6	27.4	522.0	26.7	9.4	20.9	47.5	15.3
SATS	2	78	512.0	31.4	521.3	28.1	9.3	22.1	48.7	14.8
TSSA	3	100	511.8	23.2	515.5	24.9	3.6	20.7	49.3	16.2
TASS	4	99	507.2	28.9	524.0	29.2	16.8	25.6	44.7	14.5

Note. SSTA, SATS, TSSA, and TASS describe the test and accommodation order (see Table 3). NLD = no learning disability. RLD = reading-based learning disability.

Appendix C

Repeated Measures Analysis of Variance for Each Population

Tables C1–C4 include the RM-ANOVA for each disability subgroup and grade separately.

Table C1

Repeated Measures Analysis of Variance for Grade 4 Reading-Based Learning Disability

Source	*df*	*MS*	*F*	*P*
Within subjects				
Boost	1	105,571.73	250.33***	.000
Boost x experimental group	3	105.64	0.25	.861
Error	523	421.74		
Between subjects				
Boost x experimental group	3	3,019.40	2.03	.109
Error	523	1,489.72		

*** $p < .001$.

Table C2

Repeated Measures Analysis of Variance for Grade 4 No Learning Disability

Source	*df*	*M*	*F*	*P*
Within subjects				
Boost	1	8,078.51	28.92***	.000
Boost x experimental group	3	625.76	2.24	.082
Error	650	279.39		
Between subjects				
Boost x experimental group	3	6,631.57	3.07*	.027
Error	650	2,159.48		

*$p < .05$. *** $p < .001$.

Table C3

Repeated Measures Analysis of Variance for Grade 8 Reading-Based Learning Disability

Source	*df*	*MS*	*F*	*P*
Within subjects				
Boost	1	17,801.19	70.77***	.000
Boost x experimental group	3	1,453.22	5.78**	.001
Error	372	251.55		
Between subjects				
Boost x experimental group	3	496.48	0.40	.756
Error	372	1,254.16		

**p < .01. *** p < .001.

Table C4

Repeated Measures Analysis of Variance (RM-ANOVA) for Grade 8 No Learning Disability

Source	*df*	*MS*	*F*	*P*
Within subjects				
Boost	1	904.56	4.32*	.038
Boost x experimental group	3	1,306.28	6.24***	.000
Error	467	209.22		
Between subjects				
Boost x experimental group	3	1,151.65	0.64	.589
Error	467	1,798.15		

*p < .05. *** p < .001.

Appendix D

Correlation Tables for All Subtests Administered by Group

Table D1

Intercorrelations Between Boost and Test Score (Comprehension, Fluency, and Word Recognition) for Grade 4

		1	2	3	4	5	6	7
				Student with RLD (*n* = 472)				
1	Boost (A-S)	--	-.52	.41	-.25	-.23	-.28	-.28
2	Standard GMRT		--	.56	.58	.43	.53	.50
3	Audio GMRT			--	.38	.23	.30	.27
4	WJ-Fluency				--	.66	.60	.60
5	TOSWRF					--	.53	.51
6	WJ-LWI						--	.69
7	WJ-WA							--
				Student with NLD (*n* = 604)				
1	Boost (A-S)	--	-.51	.14	-.20	-.15	-.22	-.23
2	Standard GMRT		--	.78	.60	.46	.59	.51
3	Audio GMRT			--	.55	.42	.52	.42
4	WJ-Fluency				--	.61	.54	.45
5	TOSWRF					--	.51	.40
6	WJ-LWI						--	.72
7	WJ-WA							--

Note. Sample size is slightly larger than the full sample because some students did not complete one or more of the fluency or decoding subtests. RLD = reading-based learning disability, NLD = no learning disability, A-S = audio-standard., GMRT = Gates-McGinitie Reading Tests, WJ-Fluency, WJ-LWI, and WJ-WA = the Fluency, Letter-Word Identification, and Word Attack subtests of the Woodcock-Johnson III Diagnostic Reading Battery, TOSWRF = Test of Silent Word Reading Fluency.

Table D2

Intercorrelations Between Boost and Test Score (Comprehension and Fluency) for Grade 8

		1	2	3	4	5
				Student with RLD (*n* = 373)		
1	Boost (A-S)	--	-.43	.40	-.17	-.17
2	Standard GMRT		--	.65	.47	.36
3	Audio GMRT			--	.33	.22
4	WJ-Fluency				--	.59
5	TOSWRF					--
				Student with NLD (*n* = 463)		
1	Boost (A-S)	--	-.43	.22	-.03	-.09
2	Standard GMRT		--	.79	.47	.36
3	Audio GMRT			--	.49	.32
4	WJ-Fluency				--	.54
5	TOSWRF					--

Note. Sample size is slightly larger than the full sample because some students did not complete one or both of the fluency subtests. RLD = reading-based learning disability; NLD = no learning disability; A-S = audio-standard. GMRT = Gates-McGinitie Reading Tests; WJ-Fluency = the Fluency subtest of the Woodcock-Johnson III Diagnostic Reading Battery; TOSWRF = Test of Silent Word Reading Fluency.

Appendix E

RM-ANOVA Results After Eliminating Top Performers

Tables E1 and E2 are replications of Tables 6 and 7 from the body of the text with the sample of students truncated to students who scored 45 items correct or lower on the standard administration of the GMRT. These analyses were conducted to determine if findings were consistent even after removing students who had little or no opportunity to show a performance boost from the audio accommodation due to a ceiling effect. The results are consistent with the findings reported in the body of this report.

Table E1

Repeated Measures Analysis of Variance for Grade 4 Without Top Performers

Source	*df*	*MS*	*F*	*p*
Within subjects				
Boost	1	107,257.28	347.60***	.000
Boost x RLD	1	26,722.61	86.60***	.000
Boost x experimental group	3	136.36	0.44	.723
Boost x RLD x experimental group	3	509.50	1.65	.176
Error	1,137	308.57		
Between subjects				
Boost x RLD	1	519,914.29	340.88***	.000
Boost x experimental group	3	5,874.79	3.85**	.009
Boost x RLD x experimental group	3	3,715.50	2.44	.063
Error	1,137	1,525.22		

Note. Top performers were students who scored 45 correct or lower on the standard form.

RLD = reading-based learning disability.

$^{**}p < .01$. $^{***}p < .001$.

Table E2

Repeated Measures Analysis of Variance for Grade 8 Without Top Performers (Students Who Scored 45 Correct or Lower on the Standard Form)

Source	*df*	*MS*	*F*	*p*
Within subjects				
Boost	1	17,856.39	82.69***	.000
Boost x RLD	1	3,956.53	18.32***	.000
Boost x experimental group	3	2,556.47	11.84***	.000
Boost x RLD x experimental group	3	44.42	0.21	.892
Error	810	215.95		
Between subjects				
Boost x RLD	1	462,769.23	356.46***	.000
Boost x experimental group	3	1,471.05	1.13	.335
Boost x RLD x experimental group	3	579.21	0.45	.720
Error	810	1,298.23		

Note. Top performers were students who scored 45 correct or lower on the standard form. RLD = reading-based learning disability.

*** $p < .001$.

Appendix F

Full Analyses of Variance in Alternative Models

Table F1

Summary of Regression Analysis for Standard Model Predicting Reading Comprehension for Grade 4 Students With Reading-Based Learning Disability

Variable		*B*	*SE B*	β		R^2	Change in R^2
Model 1							
	Standard GMRT	.011	.001	.460	***	.211	
Model 2							
	Standard GMRT	.006	.001	.239	***	.310	.099
	WJ-Fluency	.014	.002	.384	***		

Note. GMRT = Gates-McGinitie Reading Tests, RLD = reading-based learning disability. WJ-Fluency = the Fluency subtest of the Woodcock-Johnson III Diagnostic Reading Battery.

*** $p < .001$.

Table F2

Summary of Regression Analysis for Standard Model Predicting Reading Comprehension for Grade 4 Students Without Disabilities

Variable		*B*	*SE B*	β		R^2	Change in R^2
Model 1							
	Standard GMRT	.015	.001	.607	***	.368	
Model 2							
	Standard GMRT	.011	.001	.446	***	.414	.045
	WJ-Fluency	.010	.001	.267	***		

Note. GMRT = Gates-McGinitie Reading Tests, WJ-Fluency = the Fluency subtest of the Woodcock-Johnson III Diagnostic Reading Battery.

*** $p < .001$.

Table F3

Summary of Regression Analysis for Standard Model Predicting Reading Comprehension for Grade 8 Students With Reading-Based Learning Disabilities

Variable		*B*	*SE B*	β		R^2	Change in R^2
Model 1							
	Standard GMRT	.012	.001	.405	***	.164	
Model 2							
	Standard GMRT	.009	.002	.315	***	.195	.031
	WJ-Fluency	.005	.001	.198	***		

Note. GMRT = Gates-McGinitie Reading Tests, WJ-Fluency = the Fluency subtest of the Woodcock-Johnson III Diagnostic Reading Battery.

*** $p < .001$.

Table F4

Summary of Regression Analysis for Standard Model Predicting Reading Comprehension for Grade 8 Students Without Disabilities

Variable		*B*	*SE B*	β		R^2	Change in R^2
Model 1							
	Standard GMRT	.014	.001	.525	***	.276	
Model 2							
	Standard GMRT	.012	.001	.426	***	.310	.034
	WJ-Fluency	.004	.001	.210	***		

Note. GMRT = Gates-McGinitie Reading Tests, WJ-Fluency = the Fluency subtest of the Woodcock-Johnson III Diagnostic Reading Battery.

*** $p < .001$.

Table F5

Summary of Regression Analysis for Audio Models Predicting Reading Comprehension for Grade 4 Students With Reading-Based Learning Disabilities

Variable	*B*	*SE B*	β		R^2	Change in R^2
Model 3						
Audio GMRT	.010	.001	.368	***	.136	
Model 4						
Audio GMRT	.005	.001	.202	***	.307	.171
WJ-Fluency	.017	.001	.446	***		

Note. GMRT = Gates-McGinitie Reading Tests, WJ-Fluency = the Fluency subtest of the Woodcock-Johnson III Diagnostic Reading Battery.

*** $p < .001$.

Table F6

Summary of Regression Analysis for Audio Models Predicting Reading Comprehension for Grade 4 Students Without Disabilities

Variable	*B*	*SE B*	β		R^2	Change in R^2
Model 3						
Audio GMRT	.015	.001	.607	***	.331	
Model 4						
Audio GMRT	.012	.001	.402	***	.400	.069
WJ-Fluency	.012	.001	.314	***		

Note. GMRT = Gates-McGinitie Reading Tests, WJ-Fluency = the Fluency subtest of the Woodcock-Johnson III Diagnostic Reading Battery.

*** $p < .001$.

Table F7

Summary of Regression Analysis for Audio Models Predicting Reading Comprehension for Grade 8 Students With Reading-Based Learning Disabilities

Variable	*B*	*SE B*	β		R^2	Change in R^2
Model 3						
Audio GMRT	.012	.001	.405	***	.088	
Model 4						
Audio GMRT	.006	.002	.211	***	.156	.068
WJ-Fluency	.007	.001	.274	***		

Note. GMRT = Gates-McGinitie Reading Tests, WJ-Fluency = the Fluency subtest of the Woodcock-Johnson III Diagnostic Reading Battery.

*** $p < .001$.

Table F8

Summary of Regression Analysis for Audio Models Predicting Reading Comprehension for Grade 8 Students Without Disabilities

Variable	*B*	*SE B*	β		R^2	Change in R^2
Model 3						
Audio GMRT	.014	.001	.525	***	.294	
Model 4						
Audio GMRT	.013	.001	.449	***	.322	.028
WJ-Fluency	.004	.001	.191	***		

Note. GMRT = Gates-McGinitie Reading Tests, WJ-Fluency = the Fluency subtest of the Woodcock-Johnson III Diagnostic Reading Battery.

*** $p < .001$.